引文元数据的自动发现和标注方法研究

——以外文引文为例

姜 霖[1,2] 王东波[3]
[1](南京大学信息管理学院 南京 210023)
[2](江苏省数据工程与知识服务重点实验室 南京 210023)
[3](南京农业大学信息科学技术学院 南京 210095)

摘要: 【目的】在总结当前引文元数据抽取方法的基础上，结合语义学知识和机器学习方法，对引文元数据的自动抽取方法进行探索。【方法】实验中采用神经网络模型对人工分割过的语料进行词向量训练。利用相同类型的元数据会相对集中地出现在向量空间中某一位置的现象，通过支持向量机分类算法实现对元数据的自动归类和标注。【结果】在以外文引文数据作为测试集的实验中，本文方法取得了较高的准确率和召回率，特别是针对引文中含有多种语言和缩写的现象，具有较好的处理能力。【局限】在对于引文元数据时间内容的细粒度抽取中存在一定的局限性。【结论】实验结果表明，此方法在引文元数据的自动发现和标注上具有良好的效果，并能很大程度地提高方法的适用性和容错率。

关键词: 引文元数据 元数据抽取 机器学习 神经网络
分类号: G254

1 引 言

在科研文献中，特别是科技文献中包含大量的引文数据。引文数据不但体现了科学发展的延续性，也体现了对他人著作权的尊重和保护。引文文献一般由标题、作者、出版商、发表时间等诸多著录要素构成。在绝大多数的文档元数据标准中，引文数据都被认为是一类重要的元数据类型，在数字图书馆和语义网建设中有诸多应用。在传统的图书馆中，文献元数据信息往往需要后期的手工抽取或录入，但随着目前文献数量的激增，依靠人工抽取或录入已不太可能。此外，大量的遗留纸质文档在数字化过程中也需要自动抽取这些文档中的元数据。引文元数据的抽取是进行领域检索、引文互联分析、文章贡献评价、主题发现等研究的基础。然而由于采用的标准不一致，所以引文元数据常常具有不同的风格，如不同的语言、主题、出版物类型(如图书、期刊、会议)所采用的引文风格就不尽相同。在引文内容上，不同引文所包含的元数据个数和排列顺序也有所不同。在英文科技文献中，常见的风格就有 APA, MLA, Chicago, AMA, IEEE 和 ACM 等 6 种[1]。正是由于引文的重要性及其风格多样性，分析挖掘引文数据中所包含的信息已经成为当前信息抽取领域一项重要而又具挑战性的工作，因此本文设计了一种基于机器学习算法实现引文元数据自动抽取和标注的方法，该算法可以规避一些在人工编撰引文数据时使用模板不一致的现象，并且具有良好的跨语言平台使用效果。

2 研究综述

引文元数据的抽取作为元数据抽取的一个子任务，在计算机和图书馆等领域有着重要的研究意义，并发展演绎出多种方法。总体上，引文元数据的抽取方法可以分为三类：基于规则、基于模板和基于机器学习的方法。

通讯作者：姜霖, ORCID：0000-0003-3211-7783, E-mail：18205185622@163.com。

基于规则的方法已经被广泛应用于现实的引文抽取系统中。例如 Wei 等[2]利用逐层标注(Layer-upon-Layer Tagging)的方法抽取引文中的元数据，利用格式属性层和字典语义层的逐步标注，实现引文元数据信息的自动标注。Besagni 等[3]提出结合词性标注和规则修正来实现引文元数据的抽取和标注。李朝光等[4]则提出利用正则表达式对论文的元数据进行抽取。

基于模板的方法也是常被采用的方法之一。基于模板的方法一般会先建立模板数据库，然后通过查找和匹配模板，完成待匹配引文的抽取。Day 等[5]统计出计算机科学领域科技文献的 6 种主要参考文献格式，构建了多层知识表示框架 INFOMAP，并以此为基础，开发了基于知识的引文元数据抽取系统，其实质是基于多层模板的元数据抽取方法。Cortez 等[6]提出一种无监督的引文元数据抽取方法，利用已存在的给定领域元数据作为训练集，自动生成模板完成抽取工作。Huang 等[7]和 Chen 等[8]使用蛋白质序列表示引文字符串，将引文模板的序列表示形式存入 DNA 数据库。然后借助在 DNA 数据库中进行相似性比较的分析工具 Basic Local Alignment Search Tool(BLAST)，为待分析的引文寻找相似的 DNA，即引文模板，最后根据匹配的模板解析引文数据。

这些基于规则或模板的方法，一般具有较高的分析效率，特别是对于规则或模板所能覆盖的引文风格，具有较高的识别率。但是研究者也意识到这种基于规则或模板的方式存在缺陷，当引进新的引文风格时，需要增添规则或模板，随着引文风格的增多，规则或模板制定的负担会越来越大，造成系统冗余度提高，适用性降低。

相对基于规则和模板的方法，很多研究者选择机器学习的方法来自动地发现和标引元数据。在自然语言处理中，很多学者利用分类算法来解决文本的序列化标注问题，例如 Han 等[9]把元数据抽取看做分类问题，并将支持向量机(SVM)方法引入到元数据抽取任务中，改进了 HMM 方法在独立性假设上所带来的不足，实验取得了令人满意的结果，但该方法同时也缺失了状态转移和观察序列之间的紧密关系。此外，目前普遍使用的方法是条件随机场模型，例如 Peng 等[10]将条件随机场(CRF)方法用于引文元数据的自动抽取中，并在论文元数据抽取的公共测试集 Cora 数据集上取得良好的抽取效果。Yu 等[11]在中文科技论文数据集上测试了使用 CRF 方法抽取论文头部和引文元数据，同样取得了良好的效果。

综上所述，基于机器学习的方法能在元数据抽取上取得良好的效果，但同时也带来了人工标注、训练时间过长等额外开销。并且由于在现实中引文风格和语种的多样性，不可能穷尽所有的引文风格，特别是在文献作者人工添加引文数据时，难免会出现错误使用模板的现象，从而很大程度上降低自动识别的精度。因此从这个意义上讲，无论是手工制定的规则或模板还是通过机器学习训练出来的模板，都不具有很强的适应性。因此笔者希望通过改进机器学习算法，增加跨语种的适应性和打破使用模板的限制，从而提高自动标注的准确性和普适性。

3 引文数据自动发现、抽取和标注的关键技术

针对引文元数据抽取中存在的问题，本文提出了一种改进型的基于特征分析的引文元数据抽取方法，摆脱了传统方法依靠抽取模板的限制，并且具有跨语言平台的优势。该方法具体的技术实现路线如图 1 所示。

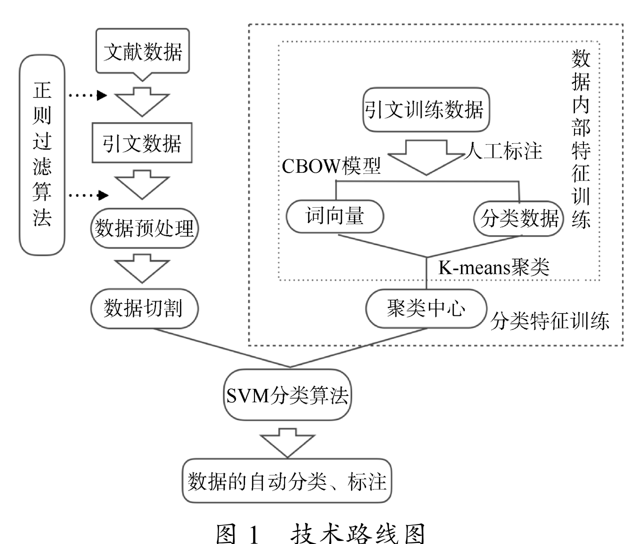


图 1 技术路线图

3.1 数据的采集以及预处理

实验使用的引文数据主要来自中文社会科学引文索引(CSSCI)引文库，由于实验中需要构建词向量空间模型，因此在处理中文引文时需要进行分词处理。分词效果会对实验结果造成较大的影响，所以在实验

中主要采用外文引文数据进行效果测试。主要通过构建正则表达式过滤引文数据，获取外文引文数据。

经过对大量外文引文数据的观察，外文引文中常以“. , : ”等符号作为元数据之间的分隔符。但很多时候“.”符号还会被用来表示人名缩写、工具的版本号等。为了提高对元数据分隔符号的辨识度，实验中制定了以下数据预处理规则。

(1) 分隔符替换规则：由于引文数据中经常出现中文标点与英文标点乱用的现象，会增加数据分隔符识别的难度，所以将全部标点都替换为英文标点。

(2) 点号替换规则：当点号前是一个大写字母且点号后是一个英文字母和标点符号时，这时往往是英文的人名；当点号前后为单个数字时，例如“Windows 3.0”，常常表示的是软件的版本号；点号还常与最近的单词组成缩写的形式，如“St.”, “Vol.”, “Aug.”等。这时将点号替换为“*”号，不再视为元数据之间的分隔符。

3.2　元数据分类特征值的训练

当前基于神经网络方法的词向量计算取得了非常好的效果，例如谷歌公司的 Mikolov 等[12]开发了一种词典和术语表的自动生成技术，能够把一种语言转变为另一种语言，取得了很好的效果。

在示例中考虑英语和西班牙语两种语言，通过训练分别得到他们的词向量空间 E(English) 和 S(Spanish)，从英语中取出 5 个词 one, two, three, four, five，设其在图 2 左部(E)中对应的词向量分别为 u1, u2, u3, u4, u5。为了方便作图，利用主成分分析(PAC)降维，得到相应的二维向量 v1, v2, v3, v4, v5。在西班牙语中取出(与 one, two, three, four, five 对应的) uno, dos, tres, cuatro, cinco，同样进行 PAC 降维处理，具体如图 2 右部(S)所示。

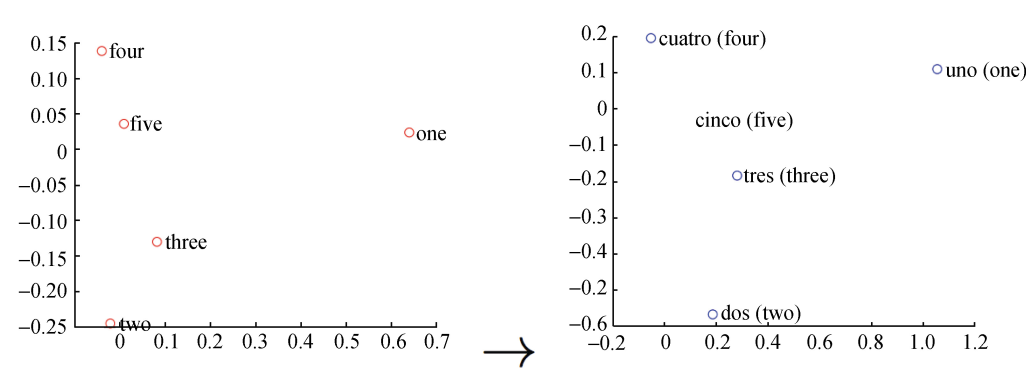


图 2　5 个词在两个向量空间中的位置

从图 2 中可以发现，5 个词在两个向量空间中的相对位置相差不多，这说明两种不同语言对应向量空间的结构之间具有相似性，从而进一步说明在词向量空间中利用距离来刻画词之间相似性的合理性，并且具有相同功能的词汇会相对集中在同一片区域。基于以上现象，利用神经网络模型为元数据中的词构建词向量空间模型，同理同一类元数据中经常出现的词会相对集中地出现在同一片区域中，利用这样的特性，笔者认为在向量空间模型中，作者名、标题名等不同类型的引文元数据会分别聚集在空间模型中的不同区域内，由此可以有效地对引文元数据实现自动标引，并且可以降低不同语言类型对于分类效果的影响。由于在中文引文中没有明显的词区分，必须借助分词软件，但是分词软件的错分现象会给实验结果带来很多干扰，因此实验中主要以外文引文为例。

在实验中，先将预处理过的训练数据进行人工识别与标注，经过标注后的数据主要有两个作用：为词向量的训练提供训练集；为 SVM 特征分析分类提供训练集。具体人工标注的样例数据如图 3 所示。

```
引文示例:
Smith , A* D*.Myth and Memories of the Nation.London:Oxford University Press.1999 : 105

训练集标注示例:
1|Smith , A* D*
2|Myth and Memories of the Nation
4|London
5|Oxford University Press
6|1999 : 105

引文示例:
Geertz , Clifford.From the Native ' s Point of View : On the Nature of Anthropological
Understanding.Reprinted in Interpretive Social Science.Berkeley:University of California
Press.1979

训练集标注示例:
1|Geertz , Clifford
2|From the Native ' s Point of View : On the Nature of Anthropological Understanding
3|Reprinted in Interpretive Social Science
4|Berkeley
5|University of California Press
6|1979
```

图 3　训练集标注示例

在图 3 中, 按照元数据的类型进行标注, 每行代表一种元数据类型。经过标注的元数据所代表的类别和表示的分类信息如表 1 所示。

表 1　训练集标注说明

分类标号	表示的分类
1	作者姓名
2	文献标题
3	期刊名或者书名
4	地点
5	出版商或者出版商
6	出版时间和页码

(1) 词向量训练

本文中单词词向量的构建主要采用 CBOW 模型[13-14], 将经过划分的元数据作为词向量构建的训练数据。该模型的主要思想是在已知当前词 W_t 的上下文 W_{t-2}, W_{t-1}, W_{t+1}, W_{t+2} 的前提下预测 W_t。图 4 给出了 CBOW 模型的网络结构, 它的结构类似神经网络模型, 主要包括三层: 输入层、投影层和输出层。

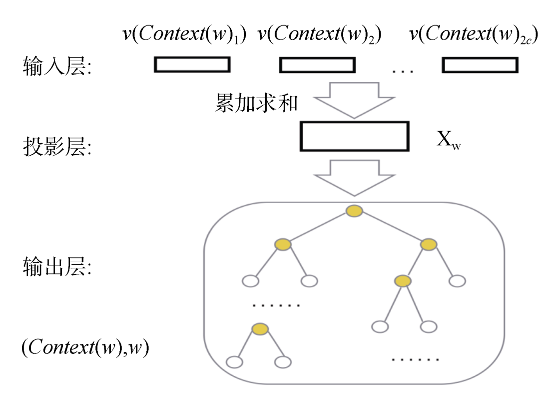


图 4　CBOW 模型网络结构

输入层: 包含 $Context(w)$ 中 $2c$ 个词的词向量 $v(Context(w)_1)$, $v(Context(w)_2)\cdots v(Context(w)_{2c})\in R^m$, m 表示词向量的长度。c 表示在词 w 的前后各取 c 个词。

投影层:将 $2c$ 个向量做求和累加, 具体如公式(1)所示。

$$X_w=\sum_{i=1}^{2c}v(Context(w_i)) \tag{1}$$

输出层: 输出层对应一棵二叉树, 它以语料库中出现的词作为叶子节点, 以各词在语料中出现的次数为权值构造出 Huffman 树, 在这棵树中叶子节点共 $N(N=|D|)$个, 分别对应词典 D 中的词。

实验中主要采用基于 Hieratchical Softmax 的 CBOW 模型。目标函数通常为公式(2)所示的对数似然函数。

$$\varsigma=\sum_{w\in c}\log p(w\,|\,Context\ (w)) \tag{2}$$

使用神经网络模型构建词向量主要有两个优势:

①词语之间的相似性可以通过词向量来体现

在神经网络概率语言模型中假定了“相似”的词对应的词向量也是相似的。并且概率函数关于词向量是光滑的, 词向量中的一个小变化对概率的影响也只是一个小变化。

②基于词向量的模型自带平滑功能, 由于 $p(w|Context(w))\in(0, 1)$, 不为零, 所以不需要额外处理。

在向量空间模型中, 作者名、标题名、期刊名、时间等不同引文分类元数据包含的词向量会分别相对集中在一个稳定的区域内, 这也使得利用分类算法对引文元数据实现自动分类和标引成为可能。

(2) 分类特征训练

由于每个类别的元数据相对集中在同一个空间区域内, 对训练数据中每个类别的词向量进行聚类计算求出聚类中心, 也就是每个分类中最具代表性的元数据, 利用待判断的词在空间模型中的位置与各个分类中心的距离, 从而判断新词的归类。在实验中对训练数据中的文本按类别进行整理, 利用词向量和 K-means 聚类算法, 分别求出每个类别的聚类中心。K-means 是一种常用的聚类算法, 对于给定的一个包含 n 个 d 维数据点的数据集 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$, 其中 $X_i\in R^d$ 以及要生成的数据子集的数目 K, K-means 聚类算法将数据对象组织为 K 个划分 $C=\{C_k,k=1,2,3\cdots\}$, 每个划分代表一个类 C_k, 每个类有一个类别中心 t_i。选取欧式距离作为相似性和距离判断准则, 计算该类内各点到聚类中心 t_i 的距离平方和。

$$J(C_k)=\sum_{x_i\in c_k}\|x_i-t_k\|^2 \tag{3}$$

聚类目标是使各聚类总的距离平方和 $J(C)=\sum_{k=1}^{k}J(C_k)$ 最小。

$$J(C)=\sum_{k=1}^{K}\sum_{x_i\in c_k}\|x_i-t_k\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}d_{ki}\|x_i-t_k\|^2$$

$$其中,\ d_{ki}=\begin{cases}1,\ 若x_i\in C_i\\0,\ 若x_i\notin C_i\end{cases} \tag{4}$$

根据最小二乘法和拉格朗日原理，聚类中心 t_k 应该取为类别 C_k 各数据点的平均值。K-means 聚类算法从一个初始的 K 类别划分开始，然后将各数据点指派到各个类别中，以减小总的距离平方和。因为 K-means 聚类算法中总的距离平方和随着类别个数 K 的增加而趋向于减小(当 $K=n$ 时，$J(C)=0$)。因此，总的距离平方和只能在某个确定的类别个数 K 下，取得最小值。利用聚类算法可以知道每类元数据中最具该类元素特征的数据占据的位置(聚类中心)，并使用聚类算法来指导元数据的分类，在分类时为数据特征降维，以缩短训练的时间。

由于在外文引文中常用“,.:”作为引文元数据的分隔符，并且每个分隔符内的数据都是同一种数据类型，每个切割部分的质心(聚类中心)到各个分类的质心的欧式距离作为分类的特征，以此对引文元数据进行分类，如图 5 所示，利用质心到质心的距离作为分类特征，可以减少分类特征数量并且强化特征描述。

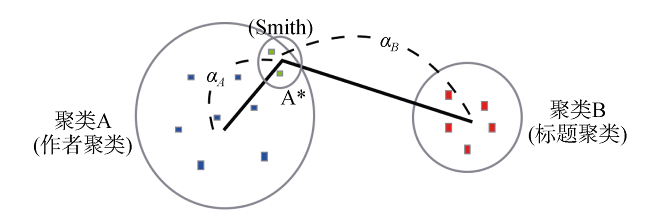


图 5　分类特征训练说明图

因为元数据中每个元素属于的类别还与在引文中的位置信息有重要的联系，所以在分类时，还以每个分割块所属的位置除以总的分割块个数得到每个分割块在引文中的相对位置，作为在分类中的位置特征值。假设切割后的引文数据表示为 $q=\{q_1,q_2,\cdots,q_i,\cdots,q_n\}$，切割单元的位置特征信息可以表示为 (i/n)。参与分类特征训练的特征采集样例如表 2 所示。

表 2　SVM 采集的特征数值样例

单元内容	离聚类 1 的距离	离聚类 2 的距离	离聚类 3 的距离	离聚类 4 的距离	离聚类 5 的距离	离聚类 6 的距离	切割单元位置特征
Chatterjee	169.70	172.06	140.57	101.79	53.43	138.36	0.17
S*	57.93	55.77	86.09	124.75	174.15	89.56	0.33
Regression and Analysis by Example	17.64	17.11	18.00	56.29	106.29	20.70	0.50
John Wiley & Sons Inc	110.96	113.44	81.81	43.00	13.34	80.03	0.67
2000	164.11	166.58	135.09	96.33	48.81	132.70	0.83
248	168.45	170.95	139.48	100.74	52.93	137.23	1.00

结合 CBOW 算法、K-means 算法以及元数据位置特征对元数据所具有的向量空间特征进行整合，使得相同类别的元数据分布在向量空间中相对集中的区域，利用分类算法进行引文元数据的自动识别和标注。

(3) 支持向量机分类

SVM 是机器学习研究中的一项重大成果，是一种重要的分类算法。它主要用于解决二值分类的模式识别问题。SVM 是在统计学习理论(Statistical Learning Theory，SLT)的基础上发展出来的一种新的通用学习方法，其核心内容是 Stitson 等[15]在 1992 年到 1995 年间提出的。采用支持向量机方法的主要优点是：

①SVM 专门针对有限样本的情况，其目标是得到现有信息下的最优解而不仅仅是样本数趋于无穷大时的最优值；

②算法最终将转化成一个二次型寻优问题，从理论上说，得到的将是全局最优点，解决了在神经网络方法中无法避免的局部极值问题；

③将实际问题通过非线性变换转换到高维的特征空间，在高维空间中构造线性判别函数来实现原空间中的非线性判别函数，特殊性质能保证机器有较好的泛化能力，同时巧妙地解决了维数问题，其算法复杂度与样本维数无关。

在综合比较神经网络模型和 SVM 模型的特点后，实验中主要选择 SVM 算法进行引文元数据特征的分类训练。对经过预处理的引文数据，按照常用的数据元分隔符进行数据切割，通过聚类算法求解当前切割元的聚类中心到各个分类的聚类中心的距离，并结合切割元所处的位置特征值作为分类特征，对切割元所属的类别进行自动分类。

4 实 验

4.1 实验评价指标

在此次实验中采用准确率和召回率以及调和平均数(F1 值)作为评价参考，公式如下。

$$准确率=\frac{提取的正确的信息条数}{提取出的信息条数}$$

$$召回率=\frac{提取出的正确的信息条数}{样本中的信息条数}$$

$$F1值=\frac{2\times准确率\times召回率}{(准确率+召回率)}$$

4.2 实验结果

以 CSSCI 采集到的 2 000 条外文引文数据为实验数据，经过人工标注后作为实验训练集。部分实验结果如图 6 所示。

```
Kumar N*.Globalization and the Quality of Foreign Direct Investment.New Delhi:Oxford University Press, 2002.72 (Nov-Dec)
.....................................................................
分类结果1.0 Kumar
分类结果1.0 N*
分类结果2.0 Globalization and the Quality of Foreign Direct Investment
分类结果4.0 New Delhi
分类结果5.0 Oxford University Press
分类结果6.0 2002
分类结果6.0 72 ( Nov - Dec )

Zhao,Quansheng.A New Era for U.S.-China Relations.Thirty Years of China-U.S. Relations: Analytical Approaches and Contemporary Issues.Lanham, Maryland: Rowman & Littlefield Publishers, Inc: 124
.....................................................................
分类结果1.0 Zhao
分类结果1.0 Quansheng
分类结果2.0 A New Era for   US  - China Relations
分类结果2.0 Thirty Years of China -   US  Relations
分类结果2.0 Analytical Approaches and Contemporary Issues
分类结果4.0 Lanham
分类结果4.0 Maryland
分类结果5.0 Rowman & Littlefield Publishers
分类结果5.0 Inc
分类结果6.0 124
```

图 6　实验结果展示

通过图 6 可以发现，实验方法对于多单位联合出版的出版社名称识别和对于使用不同引文标注风格的分割单元识别取得了良好的效果。通过语义分析可以准确地标注出时间缩写例如“Nov”和“Dec”，凸显出结合语义对引文元数据实现自动标注的手段相对于单一采用形式模板进行自动标注识别的优势，避免了当有新的模板形式添加时，需要不断地调整模板，增加程序的复杂度。此外结合语义对引文进行标注识别可以很好地规避现实中错误使用分隔符号的现象，提高了算法的容错性，增加了算法的普适性。当然这样的方法也存在一定的缺陷，比如在识别出版年份和出版页码时，由于年份和页码都是由数字组成，语义差别不大，单靠语义模型很难将元数据区分出来，如果能够配合使用模板，将会取得更好的效果。

4.3 对比实验结果分析

自然语言处理中常用隐马尔可夫模型、条件随机场模型和最大熵模型来解决序列化标注问题，当前普遍使用的模型是条件随机场模型(CRF)。为了能够可以更加凸显实验的效果，笔者使用条件随机场模型作为参照组进行了对比实验，CRF 是由 Lafferty 等[16]提出，结合了最大熵模型和隐马尔可夫模型特点的一种无向图模型，近年来在分词、词性标注和命名实体识别等序列标注任务中取得了很好的效果。

由于 CRF 实验中也需要大量的人工标注，为了减少人工标注的工作量同时考虑到两种方法对于作者姓名和数字日期、页码的识别都有很高的准确度，难以表现实验的效果，因此仅对出版社名称的抽取进行对比实验。在实验中，利用斯坦福大学自然语言研究小组推出的语法解析工具 Stanford Parser 作为英文引文的词性标注工具，在对语料进行标注时，使用五元标记模式，具体标注规则如表 3 所示。

表 3　标注规则示意表

标记符号	表示含义
B	Begin 出版社名称的开始
C	Continue 连续，名称未完结
E	End 出版社名称的结束
SW	Single Word 单个词的出版社名称
N	Not 非出版社名称词

对 2 000 篇引文数据进行人工标注，标注的具体形式如表 4 所示。

表 4　CRF 训练集的标注形式

词	词性标注	识别序列标注
Ollman	NNP	N
,	,	N
Bertell	NNP	N
Left	VBN	N
Academy	NNP	N
-	:	N
Marxist	JJ	N
Scholarship	NN	N
on	IN	N
American	JJ	N
Campuses	NNS	N
.	.	N
McGraw	NNP	B
-	:	C
Hill	NNP	C
Book	NN	C
Company	NN	E
,	,	N
1982	CD	N

具体的对比实验效果如图 7 所示。

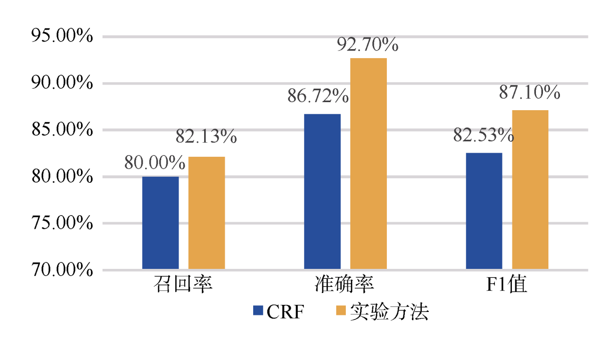


图 7　对比实验结果参数图

从图 7 中可以看出，无论是在召回率还是在准确率上，实验算法都要优于普通的 CRF 算法，特别是在识别的准确率上。由于对比实验中只选取词的词性特征作为抽取特征，这可能是导致实验效果一般的原因之一。类似 CRF 算法的经典模式识别算法，在构建模型前，一般需要事先提取特征。在提取诸多特征后，还要对这些特征进行相关性分析，找到最能代表字符的特征，去掉与分类无关和自相关的特征。因此这些特征的提取会太过依赖人的经验和主观意识，提取到特征的不同对分类性能影响很大，甚至提取特征的顺序也会影响最后的分类效果。实验算法中将词的语义特征作为分类特征，利用 SVM 算法进行元数据的自动标识，取得了一定的效果，特别是在针对英文中常出现名称缩写的问题上。实验算法利用模糊语义知识，对输入数据在空间上的扭曲具有很强的鲁棒性。

5　结　语

从实验结果可以发现，通过使用改进的引文数据元标注算法，能够较大程度地提高识别的准确度，其优势主要表现为三个方面：对输入数据的扭曲有很强的鲁棒性，例如英文缩写的识别(包括机构名称和出版商的缩写)；容错率高，即使使用错误的分隔符作为元数据的分割符，也可以通过语义无差别辨识；可移植性强，对于不同语种具有很好的适应性。这三个方面的优势使得实验方法相对于普通的机器学习的算法例如 CRF，具有明显的优势。但是该方法也存在一些不足，由于必须使用人工标注得到训练集，所以相对其他算法而言获取训练数据比较耗时，并且如果用来训练的数据量偏小，会造成使用神经网络算法构建的词向量模型不合理，进而得不到最为理想的分类效果，降低识别的准确率和召回率。如果要对引文数据实现更精确的识别，例如出版年份和页码，它们都是由数字组成，语义间的差别不大，结合模板方法可以更有效地提高识别的精度。在今后的元数据自动识别实验中构建机器学习和规则模型相结合的混合智能识别算法将可以取得更好的识别效果。

参考文献：

[1] 蒋新. 英美学术文献的几种主要引文方式[J]. 图书与情报, 2003(3): 26-30. (Jiang Xin. Several Main Quotation Ways in British-American Academic Documents [J]. Library and Information, 2003(3): 26-30.)

[2] Wei W, King I, Lee J H M. Bibliographic Attributes Extraction with Layer-upon-Layer Tagging[C]//Proceedings of the 9th International Conference on Document Analysis and Recognition. IEEE, 2007, 2: 804-808.

[3] Besagni D, Belaïd A, Benet N. A Segmentation Method for Bibliographic References by Contextual Tagging of Fields[C]//Proceedings of the 7th International Conference on Document Analysis and Recognition. IEEE, 2003: 384-388.

[4] 李朝光, 张铭, 邓志鸿, 等. 论文元数据信息的自动抽取[J]. 计算机工程与应用, 2002, 38(21): 189-191, 235. (Li Chaoguang, Zhang Ming, Deng Zhihong, et al. Automatic Metadata Extraction for Scientific Documents [J]. Computer Engineering and Applications, 2002, 38(21): 189-191, 235.)

[5] Day M Y, Tsai R T H, Sung C L, et al. Reference Metadata Extraction Using a Hierarchical Knowledge Representation Framework [J]. Decision Support Systems, 2007, 43(1): 152-167.

[6] Cortez E, da Silva A S, Gonçalves M A, et al. FLUX-CIM: Flexible Unsupervised Extraction of Citation Metadata [C]//Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries. ACM, 2007: 215-224.

[7] Huang I A, Ho J M, Kao H Y, et al. Extracting Citation Metadata from Online Publication Lists Using BLAST[C]//Proceedings of the 8th Pacific-Asia Conference, PAKDD 2004. Springer Berlin Heidelberg, 2004: 539-548.

[8] Chen C C, Yang K H, Kao H Y, et al. BibPro: A Citation Parser Based on Sequence Alignment Techniques[C]//Proceedings of the 22nd International Conference on Advanced Information Networking and Applications-Workshops (AINAW 2008). IEEE, 2008: 1175-1180.

[9] Han H, Giles C L, Manavoglu E, et al. Automatic Document

Metadata Extraction Using Support Vector Machines[C]// Proceedings of the 2003 Joint Conference on Digital Libraries. IEEE, 2003: 37-48.

[10] Peng F, McCallum A. Accurate Information Extraction from Research Papers Using Conditional Random Fields [C] // Proceedings of the Human Language Technology Conference of the North American Chapter of the Association-for-Computational-Linguistics. 2004:329-336.

[11] Yu J, Fan X. Metadata Extraction from Chinese Research Papers Based on Conditional Random Fields[C]//Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 2007, 1: 497-501.

[12] Mikolov T, Le Q V, Sutskever I. Exploiting Similarities Among Languages for Machine Translation [OL]. arXiv Preprint.arXiv:1309.4168, 2013.

[13] Mikolov T. Word2Vec Code [EB/OL]. [2015-09-18]. http:// word2vec.googlecode.com/ svn/trunk/.

[14] 周练. Word2Vec 的工作原理及应用探究[J]. 科技情报开发与经济, 2015 (2): 145-148. (Zhou Lian. Exploration of the Working Principle and Application of Word2Vec [J]. Sci-Tech Information Development & Economy, 2015 (2): 145-148.)

[15] Stitson M O, Weston J A E, et al. Theory of Support Vector Machines [R]. Technical Report, CSD-TR-96-17, London: University of London, 1996.

[16] Lafferty J, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [EB/OL]. [2016-07-15]. http://repository.upenn.edu/cis_papers/159.

作者贡献声明:

姜霖: 提出研究目标和技术路线, 完成实验编程, 论文撰写;
王东波: 训练数据的采集与整理, 完善研究方案, 论文修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 18205185622@163.com。

[1] 姜霖, 王东波. meteSplit_SVM.rar. 引文元数据自动发现和标注实验程序实现.

[2] 姜霖, 王东波. Train.rar. 训练语料.

收稿日期: 2016-08-18
收修改稿日期: 2016-11-06

Automatically Detecting and Tagging Foreign Language Citation Metadata

Jiang Lin[1, 2] Wang Dongbo[3]
[1](School of Information Management, Nanjing University, Nanjing 210023, China)
[2](Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023, China)
[3](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: [**Objective**] This paper proposes a new method to automatically extract bibliographic metadata, with the help of semantic knowledge and machine learning technologies. [**Methods**] We used the neural network model to create word vectors from manually split data, and then found that same type of metadata is relatively concentrated at certain locations in the vector space. Thus, we proposed a new SVM classification algorithm to classify and annotate the bibliographic metadata automatically. [**Results**] The proposed method achieved high recall and precision rates with citation data, especially for citations with various languages and abbreviations. [**Limitations**] The fine-grained extraction of the time related content could be improved. [**Conclusions**] The proposed method could effectively detect and tag bibliographic metadata, and improve the system's compatibility and fault tolerance ability.

Keywords: Bibliographic Metadata Metadata Extraction Machine Learning Neural Network